A NEW NON-PARAMETRIC METHOD FOR ESTIMATING MISSING VALUE IN MICROARRAY DATA BY GO CLUSTERING

S. F. Molaeezadeh[1], M. H. Moradi[2]
[1,2]Biomedical Signal Processing Laboratory, Faculty of Biomedical Engineering, Amirkabir University of Technology (Polytechnic), Tehran, Iran
e-mail: fmolaee@cic.aut.ac.ir, mhmoradi@aut.ac.ir

Abstract-**Microarray data often contain missing expression values. Performance of many analysis methods degrade for this problem. In this paper, an estimation method based on Gene Ontology Clustering is proposed. For missing value estimation we used biological information to select similar genes, while other methods such as k-NN used statistical criteria e.g. Pearson correlation and Euclidean distance for finding similar genes. After that, missing value is estimated by linear combination of similar genes. Experimental results on Alon colon dataset show two benefits. First, our method has better performance than the other imputation such as k-nearest neighbor (k-NNimpute) and Local Least Squares imputation (LLSimpute). Second, it does not need any parameter.**

Keywords-**Gene Ontology (GO), Normalized Root Mean Squared Error (NRMSE), K-nearest neighbor (k-NN), Local Least Squares (LLS)**

I. INTRODUCTION

The DNA microarray technology enables monitoring and measurement of expression levels for thousand of genes simultaneously through the hybridization process [1]. It can be applied to a wide range of studies including gene regulation, disease diagnosis and prognosis, cancer classification, bio-marker discovery and drug development [2]. A typical DNA microarray study utilizes several DNA microarray chips on different tissue samples and generates a numerical array with thousands of rows (genes) and tens of columns (experiments/DNAchips) [3]. Usually every experiment (column) contains some missing value and more than 90% of genes (rows) are affected [4].

Missing values occurs because of different reasons, for example insufficient resolution, image corruption, dust or scratches on the slides [5], blemishes on the chips [4], spotting and hybridization errors [6]. This problem degrades performance of many algorithms such as clustering, principle component analysis and singular value decomposition [5]. There are some solutions for this problem. First is to repeat the experiment but it is costly or time consuming. Second is to remove the samples that have missing values but it is not reasonable because of losing information and a limited number of microarray chips. Third is to estimate missing value. Many algorithms have been developed to estimate missing value such as k-nearest neighbor imputation and singular value decomposition based imputation [5], Gaussian mixture clustering imputation [4], Local Least squares imputation [7] and Support Vector Regression imputation [8]. Influence of both Biological Processes (BP) and Molocular Function (MF) ontologies on the imputation accuracy is investigated in [6].

In common estimation methods, similar genes are empirically determined based on statistical correlations such as Euclidean distance or by automatic k-value estimator, but these genes do not necessarily correlate biologically. So we proposed to use biological correlations for finding neighbor genes. Gene Ontology (GO) is a tool that provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes. Our method includes two parts. First, genes that are similar biologically are selected by using GO clustering. Then a gene that has missing value is estimated by linear combination of co-cluster genes. We used Part of Alon colon dataset that its GO information is available in [9] as our dataset. Test data is prepared by randomly removing 1-10% values of used data. Our method is compared with k-nearest neighbor (k-NN) [4] and Local Least Squares imputation (LLS) [7]. Criterion of comparison is Normalized Root Mean Squared Error (NRMSE). Experimental results show two benefits. First, our method has better performance than the other imputation such as k-nearest neighbor (k-NN) and Local Least Squares imputation (LLS). Second, it is a non-parametric method.

This paper is organized as follows. In section II Gene Ontology is presented. Section III describes used methods. Results are reported in section IV. We conclude this paper and give direction of future work in Section V.

II. GENE ONTOLOGY

The high throughput of recent gene expression measurement technologies like microarrays has driven the development of tools to help in the task of representing and processing information about genes, their products and their functions. One of the most important of these tools is the Gene Ontology (GO), which is being developed in tandem with work on a variety of bioinformatics databases [10]. It provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes. In 1998 Gene Ontology project initiated in order to provide a common reference framework for the associated controlled vocabularies.

Gene Ontology has two important benefits in microarray studying. First, the significantly differentiated genes from statistical analysis can be annotated with GO terms; second, the microarray data can be grouped according to the functions of the genes or biological processes they are involved. According to mentioned benefits, they reveal two points. First point is where the gene products are, what they are doing and in which biological process. The second point is very important in a sense that the information is organized in a meaningful way. We can gain a better understanding of the data than purely statistical analysis because biological significance does not necessarily have to be statistically significant. For example, for the genes involved in a biological process, they may not be significant from statistical analysis, but if they consistently change, even though in small scales, between cancer and normal tissues,

the process may be important in the understanding of the cancer [2]. It is available in [11].

III. METHODS

In this section we describe three data imputation methods: k-nearest neighbor imputation (KNNimpute), local least squares imputation (LLSimpute) and a new method based on GO Clustering (GOC). These methods are briefly introduced in this section.

A. KNNimpute

Microarry data can be represented as an M×N matrix $G = (g_{ij})_{i,j=1}^{M,N}$, where M and N are number of genes and different conditions respectively. We supposed that the missing value occurs in the first component of gene r. let C is the complete rows of G. KNN algorithm find K rows, $R_1,...,R_K$, in C that have the shortest Euclidean distances to g_r. Let d_i is the Euclidean distance between vectors $w = (g_{rj})_{j=2}^{N}$ and $w_i = (R_i)_{i=1}^{K}$. Also $B = (b_i)_{i=1}^{K}$ denote the k-dimensional vector consisting of first components of such K neighbor genes. g_{r1} is estimated as follows:

$$\hat{g}_{r1} = \frac{\sum_{i=1}^{K} b_i / d_i}{\sum_{i=1}^{K} 1/d_i} \qquad (1)$$

B. LLSimpute

We used LLS algorithm that is proposed in [7]. This method has two stages. First K similar genes are selected by KNN algorithm. Then missing value is predicted by the least squares formulation as (2). g_{r1} is estimated by

$$\hat{g}_{r1} = B^T \, pinv(A^T) \; w \qquad (2)$$

where $A = (R_{ij})_{i,j=1}^{K,N-1}$ denote K× (N-1) matrix consisting of K neighbor genes. B and w are introduced in part A and pinv is pseudoinverse.

C. GOCimpute

We collected GO annotation for genes used in Alon colon cancer microarray experiment from SOURCE [9] online database on 5/2/2006. Some genes in the original dataset do not have GO annotation in [9]. But majority of genes have been annotated. For this dataset, we found 1597 original genes, out of 2000, that were annotated with at least one of the 10847 GO annotations. Genes that have GO is used as original data. We find 1562 different GO terms in 1597 genes. Similarity criteria in our clustering method are to have same GO term. Because there are 1562 type GO terms, the number of our cluster are 1562. After GO clustering we can obtain genes belong to each GO.

Our proposed method has two parts: selecting similar genes via GO clustering and estimating missing value with using similar genes.

1) S*electing similar genes:* We get all of GOs belong to gene r and then genes belong to these GOs. These genes are similar genes to gene r.

2) Estimation: after determination of similar genes, g_{r1} is estimated by linear combination of these genes. We used (2) for predicting missing value.

D. Imputation Error Estimation

We delete between 1 and 10% of data at random to create test data. The used criterion to assess the accuracy of estimation is calculated as:

$$NRMSE = \frac{RMS(M - M_{est})}{RMS(M)} \qquad (3)$$

where M is the original data matrix and M_{est} is the estimated matrix.

IV. RESULTS

In this section, we represent results of our work. We performed mentioned estimation methods on part of colon cancer dataset that have GO annotations. This dataset is obtainable in [12]. Details of the dataset are described in Table I. We implemented mentioned methods for two conditions: the different number of similar genes (K=10, 50 and 100) and different percentages of missing values (1, 5 and 10%). We repeat estimation procedures ten times for avoiding of random result because rows (genes) that have missing value are selected randomly. Then we average calculated NRMSEs for ten times that the experiment is repeated.

TABLE I
Details of colon cancer dataset

data set	Number of classes	Number of samples	gene expression levels
Colon cancer (Alon et al., 1999)	2	62	2000

In Table II, we evaluated GOC, KNN and LLS for three k-values i.e. 10, 50 and 100. Also percentage of missing value is 1%. Results show that GOC have good performance than other methods. It does not need parameter k while result of two other methods depends on parameter k and they have bad result for larger k-values. In fig. 1, we compared three methods for different k-values.

TABLE II
NRMS error for 1% missing value

method	K=10	K=50	K=100
GOC	0.0006929	0.0006929	0.0006929
LLS	0.0007037	0.0014	0.00081532
k-NN	0.00077757	0.00081676	0.00088833

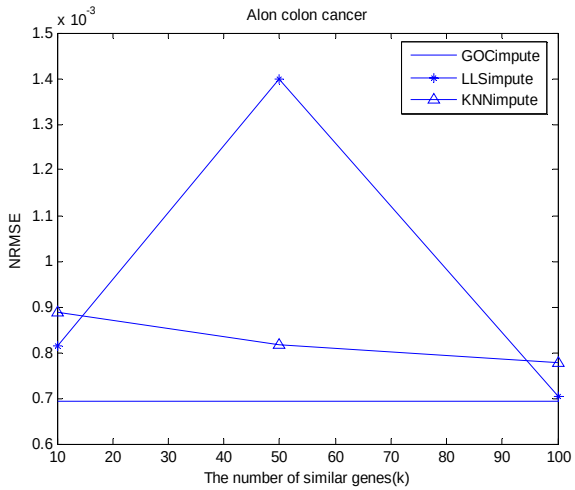

Fig. 1 comparison of three used method for different k-values and 1% missing value

In Table III, we performed three methods with fixed k-value (k=10) and different percentages of missing values (1, 5 and 10%). Results show that performance of three method are almost same for k=10.

TABLE III
NRMS error for three imputation method

Percentage of missing value	GOC	LLS (k=10)	KNN (k=10)
1%	0.0006929	0.0007037	0.00077757
5%	0.0018	0.0018	0.0018
10%	0.0025	0.0025	0.0026

V. CONCLUSION AND FUTURE WORK

We propose new method for estimating missing values based on GO clustering. Results show that this method can estimate missing value better than KNN and LLS methods. In addition our method has these benefits: the number of similar genes is not fixed, but in other methods, number of neighbor genes (k) is fixed. Also, our method is non-parametric method while it is necessary to determine k value in other methods. Parameter k is determined empirically or by automatic k-value estimator. Other difference of our method is criterion of selecting similar genes. Our criterion is based on prior biological knowledge but other criteria are based on statistical relations. We propose that biological information should be used in statistical processes related to analysis microarray data.

ACKNOWLEDGMENT

This work is supported by Iran Telecommunication Research Center (**ITRC**).

REFERENCES

[1] M. Ng and L. Chan, “Informative Gene Discovery for Cancer Classification from Microarray Expression Data,” IEEE, 2005, pp.393-398.
[2] S. Li, M. J. Becich, J. Gilbertson, “Microarray Data Mining Using Gene Ontology,” MEDINFO 2004, pp.778-782.
[3] X. Xu, A. Zhang, “Selecting Informative Genes from Microarray Dataset by Incorporating Gene Ontology,” Proceedings of the 5th IEEE Symposium on Bioinformatics and Bioengineering (BIBE’05), 2005.
[4] M. Ouyang, W. J. Welsh, P. Georgopoulos, “Gaussian mixture clustering and imputation of microarray data”, Bioinformatics, Vol. 20 no. 6 2004, pp.917-923.
[5] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Boststein and R. B. Altman, “Missing value estimation methods for DNA microarrays”, Bioinformatics, Vol. 17 no.6 2001, pp. 520-525.
[6] J. Tuikkala1, L. Elo, O. S. Nevalainen1, T. Aittokallio, “Improving missing value estimation in microarray data with gene ontology”, Bioinformatics, Vol. 22 no.5 2006, pp. 566-572.
[7] H. Kim, G. H. Golub, H. Park, “Missing value estimation for DNA microarray gene expression data: local least squares imputation”, Bioinformatics, Vol. 21 no.2 2005, pp. 187-198.
[8] X. Wang, A. Li, Z. Jiang, H. Feng, “Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme”, BMC Bioinformatics, 2006, pp. 1-10.
[9] http://smd.stanford.edu/cgi-bin/source/sourceBatch Search
[10] B. Smith, J. Williams and S. Schulze-Kremer, “The Ontology of the Gene Ontology,” Proceedings of AMIA Symposium 2003.
[11] http://www.geneontology.org
[12] Colorectal Cancer Microarray Research [http:// microarray. princeton.edu/ oncology/].